Chapter 7

DISCUSSION PAPER ON SPEECH PERCEPTION

Wayne A. Wickelgren

Department of Psychology
University of Oregon

My comments on Dr. Lehist's paper are organized primarily into two sections: (a) serial vs. parallel processing in speech recognition. First, I will argue that speech production involves a special kind of mixture of both serial and parallel processing, though the observable output of the speech production process makes the serial aspect of speech production much more obvious than the parallel aspect. Second, I will argue that speech recognition also involves a special kind of mixture of serial and parallel processing, but with parallel processing overwhelmingly more important at linguistic levels below the word.

Throughout this paper I will be defending the position that the context-sensitive allophone representative is the important unit at the segmental level in both speech recognition and articulation. In my opinion, context-sensitive allophones subserve the proposed functions of both syllables and phonemes better than do syllables and phonemes, at the segmental level in speech perception and production. However, I strongly suspect that representatives of context-free phonemes play an important role in the child's acquisition of context-sensitive allophones in articulation. Furthermore, the concept of the phoneme plays an important role in the definition of the context-sensitive allophone.

I assume that the units of representation in the lexicon (dictionary) of a language are concepts, not

words or morphemes. By this I mean that every different meaning of a word will have a different unit representing it in the lexicon. In addition, a word made up of two morphemes such as "blackbird" will often not be represented by two morpheme units, "black" and "bird", but rather by one concept unit, "blckbird". However, in some cases, such as the representation of singular vs. plural forms of a concept, I would assume that a plural concept such as "birds" is represented by the conjunction of two concept representatives, "bird" and "plural". A regular past tense verb, such as "walked" is probably represented by "walk" and "past", etc. In addition, I assume that two "synonymous" meanings of different words have at least somewhat different representation at the concept level, though all semantically similar concepts presumably have strong associations between them or overlapping "semantic feature" representation in the concept system, etc.

To communicate concepts by speech, we must translate concepts into a sequence of articulatory gestures which result in a temporally distributed speech wave form. For the moment, let us ignore the difficulty involved in segmenting speech at any articulatory, acoustic, or auditory-sensory level, and make the assumption that a concept (word) is represented by a set of structural (articulatory and auditory) units at some lower phonetic level(s) of the nervous system.

In the past, linguistics has always considered the structural (phonetic) analysis of a concept (word) to be an ordered set of nonoverlapping segments. Occasionally, it has been argued that the immediate segmental constituents of words are syllables, but more frequently the assumption has been that the immediate constituents of words are phonemes. In either case, the representation of a word is by an ordered set of "context-free" segments. By "context-free", I mean that the same segments can appear in a large variety of different segmental contexts. That is to say, the /s/ phoneme in the word "struck"

is the same /s/ phoneme as in the words "pass" or "pensive". The word must be represented by an ordered set of such segments (phonemes), because, frequently, the same segments (phonemes) in a different order represent a different word. Of course, there are a number of phonological restrictions and differential statistical probabilities of one segment (phoneme) following another segment (phoneme) in the language, but this fact is not represented in the phonetic "spelling" of any word in the lexicon.

Recently, Wickelgren (1969(a) and (b)) proposed a rather different type of immediate constituent analysis of concepts (words) at a "segmental" level. Wickelgren proposed that a word like "struck" be represented by an unordered set of context-sensitive allophone representatives such that each allophone representative was essentially an ordered triple of immediately adjacent phonemes in the phonemic spelling of the word. Thus, the spelling of the immediate constituents of the word "struck" would be
$\#^{s}t$, $s^{t}r$, $t^{r}\Lambda$, $r^{\Lambda}k$, $\Lambda^{k}\#$.

For convenience, the context-sensitive allophonic spelling of the word "struck" has been written in the obvious order. However, the representation of "struck" can be by the unordered set of context-sensitive allophones at some level of the nervous system, since the order of the context-sensitive allophones can be uniquely reconstructed from associations in long-term memory. Clearly, the association from $\#^{s}t$ will be strongest to $s^{t}r$ of all the context-sensitive allophones in the word, and the association from $s^{t}r$ will be strongest to $t^{r}\Lambda$ of all the allophones in the word etc. Thus, a simple left-to-right associative generation process will reconstruct the order of these context-sensitive allophones from the unordered set. Note that the immediate phonetic constituents of words by this theory are context-sensitive. The selection of one constituent for a word places restrictions on what other constituents can be selected in the spelling of the word. Another way to say this is that the immediate phonetic constituents are

overlapping, rather than non-overlapping as in the case of phonemic or syllabic spelling. Although a context-sensitive allophonic spelling of a word does dictate some particular order in the articulation of the phonetic constituents, these allophonic constituents overlap to a certain limited degree (in terms of phonemes, each constituent overlaps its two immediate left and right neighbors). Thus, the order of the constituents is not like the segments of a tape so much as it is like the links of a chain, each of which interlocks with two adjacent links. Another useful analogy is that the context-sensitive allophone is like a piece in a linear jigsaw puzzle with notches and tabs that exactly fit the tabs and notches on the correct left and right hand pieces. It may be somewhat missleading to refer to context-sensitive allophones as segmental representatives at all. Perhaps context-sensitive allophones should be called links or "linkments" (the latter by analogy to "segment"). I will not pursue further this attempt to add another word to the English lexicon.

Presumably, at still more peripheral levels of the nervous system, a particular context-sensitive allophone activates a particular set of motor feature representatives that control the muscles of the vocal tract. Also, at a peripheral auditory level, sets of auditory feature representatives are associated with each context-sensitive allophone for the purposes of speech recognition. I will have very little to say regarding either the auditory or articulatory feature levels of the speech articulation and perception processes in this paper.

In addition, I will have nothing to say in this paper concerning historical linguistics and phonology. Although it is not necessarily true, it seems very plausible to me to assume (as I do) that the processes governing changes in the sound system of a language are very different from the processes controlling performance in speech recognition and articulation by a competent adult speaker. In agreement with Ladefoged (1970), I believe that a large number of phono-

logical laws are essentially laws of historical linguistics, not rules that function in the recognition or production of speech. However, also in agreement with Ladefoged, I suspect that some phonological rules (e.g., regular plural formation, regular past tense formation) are rules which do function in adult speech production and recognition processes. Unfortunately, I have not given sufficient time to these matters to make discussion in the present paper worthwhile. In any event, the casual observation by Lehiste of nonadjacent contextual influences in historical linguistics seems to me to be not clearly relevant to the possible role of context-sensitive allophonic coding in adult speech performance.

SERIAL AND PARALLEL PROCESSES IN SPEECH PRODUCTION

Lehiste cites Öhman (1966) who showed in VCV sequences that the transition from the first vowel to the following consonant depended on the nature of the second vowel. In agreement with the terminology of Daniloff and Moll (1968), let us call this an example of "forward" co-articulation. Öhman (1966) also demonstrated that the transition from the intervocalic consonant to the second vowel depended upon the nature of the first vowel. Following the terminology of Daniloff and Moll, this is an example of "backward" co-articulation. It is almost universally assumed that backward co-articulation effects can be explained entirely by mechanical inertial factors in which changes in the state of contraction of articulatory muscles lag behind the arrival and termination of neural commands to an articulatory muscle persist through subsequent segments unless directly contradicted by a command to an articulatory muscle.

Forward co-articulatory effects cannot be explained by assuming that the neural commands for each segment are not delivered strictly in succession, but rather are delivered in an overlapping (shingled) manner to the articulatory muscles. A somewhat vaguer "explanation" of forward co-articulation is

that there is forward "planning" of a larger portion of the utterance than a single segment.

In my opinion, the most attractive explanation of at least some forward co-articulatory effects is the "priming" mechanism suggested by Lashley (1951), in which all of the segments in a phrase are partially activated (primed) before beginning to fully activate any single segmental representative. Lashley considered this priming process to be necessary in order to account for anticipatory errors in pronunciation. Wickelgren (1969(a) and (b)) points out how the priming process, in conjunction with the assumption that the segments are context-sensitive allophones, provides a mechanism for articulation of an entire phrase as an automatic process at a "lower" phonetic level without continued direction by the higher cognitive (syntactic and semantic) level. This would permit the cognitive level to be planning the next phrase while the phonetic level of the nervous system was directing the articulation of the previously planned phrase. MacKay (1969, 1970, 1971) has more fully developed the necessary characteristics for this priming process and carefully documented the power of this priming process in explaining a variety of speech errors. MacKay's error analyses indicate the need to assume a temporal gradient of priming with the greatest degree of priming being for the next segment to be uttered, the next greatest degree for the following segment, etc.

MacKay assumes that this gradient of priming is achieved by a scanner passing over representatives of successive segments arranged in a non-associative buffer memory. However the associative-chain theory proposed by Wickelgren (1969(a) and (b)) for speech articulation provides a completely natural mechanism for achieving exactly this type of priming gradient. Consider the associative chain of five consecutive context-sensitive allophones for the word "struck" as shown in Fig. 1. During articulation of $_{s}tr$, the representative of $_{s}tr$ will be maximally activated, but it will be sending impulses "downstream" in the

chain to further increase the degree of activation of $_{+}r_{\Lambda}$ above the level produced by the prior phrase priming process (selection of the unordered set of context-sensitive allophones). This heightened activation (priming) of $_{+}r_{\Lambda}$ should result in some degree of heightened priming of $_{r}\Lambda_{k}$, and so on to the end of the associative-chain. This would provide precisely the gradient of activation that MacKay's analysis indicates is necessary in order to account for the distribution of a variety of speech errors as a function of segmental (phonemic or allophonic) distance.

An illustration of the qualitative characteristics of this priming process for each of the five context-sensitive allophone representatives in the word "struck" as a function of time is shown in Fig. 2. Note that in Fig. 2 all of the context-sensitive allophones in the unordered set for the word "struck" start off at a positive level of priming due to the phrase priming process that occurred when word representatives at the concept level selected the appropriate unordered sets of context-sensitive allophone representatives at the segmental level. In turn, each of the successive context-sensitive allophone representatives in the word is raised to a level of maximum activation. At any given point in time, the degree of activation is maximal for the current context-sensitive allophone representative, next highest for the immediately following context-sensitive allophone representative, next highest for the following allophone representative, etc. Allophone representatives that are not a part of the unordered set for any word to be articulated in the phrase are at the lowest level of activation of all, arbitrarily called zero level of activation. After a previously-primed allophone representative has been articulated it is inhibited and returned to zero degree of activation, unless that allophone representative has been "doubly primed" (primed twice in conjunction with two occurrences in the phrase). In this latter case, it is necessary to assume the al-

lophone representative returns to the level appropriate for a singly-primed allophone representative. For phrases of any reasonable length in English, this repetition of an allophone representative in a phrase will occur only very rarely.

An extremely interesting experiment by Ladefoged and Silverstein (1970) on the speed with which a subject can interrupt a currently-articulated utterance to begin a new utterance provides evidence for precisely the type of phrase priming postulated by Wickelgren (1969(a) and (b)). Ladefoged and Silverstein found that there were no differences in the speed with which subjects were able to stop saying what they intended to say and start saying something else as a function of where the cue to do this was given, provided it was given during the utterance. That is to say, there were apparently no stress-linked or syntax-linked differences in ease of interrupting speech during the articulation of a phrase. However, during a period just prior to the subject's beginning articulation of the utterance, the subject responded much more slowly to the cue (as long as 750 msec before the utterance vs. an average of 350 msec during the utterance). Ladefoged and Silverstein interpreted their results to indicate that, prior to beginning an utterance, the speaker was planning the articulation and could not readily plan another utterance. By this same token, it must be assumed that during articulation of the utterance, this higher syntactic and semantic (concept) planning level is free to plan the articulation of another utterance and not involved at all in the ongoing control of articulation of the current utterance. This latter assumption accounts for the absence of any syntax-linked differences in ease of interruption during the utterance.

Priming (or some process like priming) seems necessary in order to account for the more remote forward co-articulatory effects such as described by Daniloff and Moll (1968). However, with the assumption of context-sensitive allophonic coding in speech articulation, priming is not necessary in order to

account for the more remote forward co-articulatory effects such as described by Daniloff and Moll (1968). However, with the assumption of context-sensitive allophonic coding in speech articulation, priming is not necessary in order to account for the more immediate forward co-articulatory effects such as those observed by Öhman (1966). In VCV utterances, Öhman apparently found no effects of the first vowel on the steady state formant levels for the terminal vowel and no effects of the terminal vowel on the steady state formant levels for the initial vowel. Thus, all of the co-articulatory effects observed by Öhman can be characterized as being effects of the initial and terminal vowels on the intervocalic consonant. The capacity for such immediate context-conditioned variation is an obvious consequence of the assumption that the segmental units are context-sensitive allophone representatives, rather than phoneme representatives. Clearly, the context-sensitive allophone representative for ${}_{\alpha}b_{u}$ will be different from that for the context-sensitive allophone ${}_{o}b_{u}$. Since the segmental representatives of the consonant are different in these two cases, the consonant can be different in both its initial and terminal transitions. There is no necessary reason within context-sensitive coding theory to assume that the only effect of a prior segment would be upon the initial portion of the subsequent segment, though it is reasonable to assume that co-articulatory effects will have some gradient of this type.

According to the theory developed in this section, speech articulation involves a combination of serial and parallel processes. Phrase priming may be an entirely parallel process, that is to say, all of the allophones for all of the words in the phrase may be primed simultaneously. Alternatively, as suggested by Wickelgren (1969(a)), each word in the phrase may be primed in its appropriate temporal order. However, the priming of the segments of each word consists of the simultaneous priming of all of the allophones in the word (parallel process). In

any event, the maintenance of this priming for all of the allophone representatives in the entire phrase during articulation of the phrase is assumed to be a parallel process. The succession of maximally activated allophone representatives clearly has the primary character of a serial process, but the existence of a priming gradient induced by the associative-chain is once again a parallel process. The advance priming of immediately succeeding allophone representatives provides an extremely natural mechanism, within an associative-chain theory, by which a certain degree of temporal overlap in the neural commands for successive segments might be achieved.

SERIAL AND PARALLEL PROCESSES IN SPEECH RECOGNITION

Correlation of recognition of different segments.

Lehiste raises an interesting criticism of the context-sensitive coding theory in speech recognition based on the data of Lehiste and Shockey (1971). Lehiste argues as follows:

> "It seems reasonable to assume that if the context-sensitive allophone is the minimal unit of perception, the context to which the allophone is sensitive should be perceptible. Thus, the /p/ should be equally perceptible, i.e. equally recoverable, under all three conditions described above." (These three conditions are /api/, /apa/, and /ap#/.)

Lehiste and Shockey (1971) investigated the identifiability of each vowel and consonant segment in a V_1CV_2 sequence both when the sequences were intact and when either the first or second half of the utterance was removed by cutting the tape during the voiceless plosive gap (the consonants were either /p/, /t/, or /k/).

To examine the validity of Lehiste's assertions regarding context-sensitive coding in perception, it is necessary to discuss the presumed operation of context-sensitive allophones in speech recognition, in

some detail.

From a perceptual point of view, the context-sensitive coding theory asserts that the acoustic features which contribute to the recognition of adjacent context-sensitive allophones must be to some extent overlapping in time of occurrence. To illustrate this, consider the hypothetical representation shown in Fig. 3.

Fig. 3 illustrates one possible set of distributions for the density of features for each allophone in the word "struck". The feature densities shown in Fig. 3 are surely incorrect in a number of respects. First, since by various approximate measures of phoneme duration at a peripheral articulatory and acoustic level, the different types of phonemes differ in their relative durations, it is likely that the spread of the distributions of features for different context-sensitive allophones would have to be assumed to be somewhat different as well. Fig. 3 shows all of these spreads to be approximately equal, and this is probably false. In addition the unimodal "normal-type" distributions shown in Fig. 3 are just a wild guess as to the approximate form of the distributions. Nevertheless, for present purposes, the overlapping feature-density distribution shown in Fig. 3 are completely satisfactory.

Note that in Fig. 3 the cues for recognizing the allophone /$_{s}$t$_{r}$/ occur at the same time as many of the cues appropriate for recognizing the allophone /#st and the allophone /$_{t}$r$_{\Lambda}$/. In Fig. 3, there is even some temporal overlap between the features for /$_{s}$t$_{r}$/ and the features for /$_{r}\Lambda_{k}$/, though the decision to represent the spreads in this manner was made purely arbitrarily to illustrate the possiblity of some more remote interaction that is nevertheless consistent with the formulation of context-sensitive allophones as the basic units in perception. It would be simpler to assume that non-adjacent allophones had no temporal overlap in their features, but I have no way of knowing that this is true at the present time. The point is that the representation shown in Fig. 3 is

perfectly consistent with the basic idea of context-sensitive coding in terms of phoneme triples, and yet it does yield some temporal overlap in the features for allophones separated by one intervening allophone. However, I think that it is contradictory to my context-sesitive coding theory to have temporal overlap in the features for allophones separated by two or more intervening allophones.

One should be careful to note that the existence of some temporal overlap between two adjacent or non-adjacent context-sensitive allophone representatives does not imply that the features that contribute to the recognition of each allophone during this region of temporal overlap are the same. Indeed, these features may have nothing in common whatsoever. Chances are, considering what we know about "transitions" between successive phonemes (allophones) that the cues for immediately adjacent context-sensitive phonemes (allophones) do have much in common in addition to their time of occurrence. Undoubtedly, it is often the case that many of the same features contribute to the recognition of immediately adjacent context-sensitive allophones.

Presumably, the features that contribute to the recognition of any given context-sensitive allophone are somewhat redundant. That is to say, one can fail to perceive some of these features (as a result of either external or internal noise) and still be able to activate the correct context-sensitive allophone representative. This somewhat complicates the interpretation of any experiment in which the recognizability of an allophone (phoneme) was investigated as a function of cutting a tape recording at various points or adding different types of noise, etc. Certainly, one cannot assume that eliminating any particular time segment during the region of positive feature density for any particular allophone would necessarily reduce recognition of that allophone, especially under conditions that otherwise produce very high intelligibility for the allophone.

Excluding the possibility of attentional fluctuations or other confounding factors in speech recognition, context-sensitive coding makes the prediction that the recognizability of adjacent phonemes should be positively correlated. The cues for adjacent phonemes undoubtedly have much in common, at a minimum they have time of occurrence in common. Thus, many factors that influence the recognizability of one allophone must also affect the recognizability of an immediately adjacent allophone. If zero or negative correlation is found for the recognizabilities of immediately adjacent phonemes, it would be a serious disconfirmation of context-sensitive coding in speech perception.

Looking at it from a somewhat different point of view, if one has activated the internal representatives (recognized) /$_{s}t_{r}$/ then one ought, logically, to know that the immediately prior phoneme was /s/ and the immediate subsequent phoneme was /r/. One would not know from this what the exact context-sensitive allophones were for the immediately prior and succeeding segments, but one ought to be able to write down a all three phonemes, given only the recognition of the medial context-sensitive allophone. Of course, it is possible that although this information is logically present in the nervous system, people do not make use of it in speech recognition, but this seems extremely unlikely.

However, this raises the point that, because one has recognized a single phoneme from an utterance of several phonemes, one cannot assume (in fact it would be unreasonable to assume) that this has occurred because the individual has activated the particular context-sensitive allophone representative appropriate for that phoneme. Presumably, if a subject can only identify a single phoneme from some utterance, he has not maximally activated any single context-sensitive allophone representative consistent with that phoneme. Rather he has activated a set of context-sensitive allophones appropriate for that phoneme, but no single member of this set (no single context-sensitive allo-

phone representative) has been activated more than the others. This would permit him to say that a particular phoneme had occurred, but not to say what the immediately prior or succeeding phonemes were.

In light of the above discussion, we are now in a position to evaluate the validity of Lehiste's comments concerning the significance of the Lehiste and Shockey experiment for context-sensitive coding theory. For the purposes of discussing the Lehiste and Shockey experiment I will use an illustration similar to that of Fig. 3 for one of the V_1CV_2 triples used by Lehiste and Shockey. Fig. 4 illustrates the approximate feature density for each consecutive context-sensitive allophone in the V_1CV_2 utterance /api/. The dotted vertical line illustrates the approximate position of the plosive break at which Lehiste and Shockey made cuts in the tape under conditions for presenting either /ap/ or /pi/.

As Fig. 4 illustrates, presenting only /ap/ from the triple /api/ by means of a cut at the plosive break might have very little effect on the recognizability of the /a/ phoneme. This is because very little of the features necessary for the recognition of /a/ occur to the right of the plosive break. However, these features may be critical for the identifiability of /#ap/ as a particular context-sensitive allophone. Furthermore, the degree of overlap between successive allophones shown in Fig. 4 is a wild guess. The percentage of features for /#ap/ to the right of the plosive break may be considerably greater than that shown in Fig. 4.

Presumably in those cases when subjects were only able to identify the initial vowel of such an utterance, the cut did interfere with some features that were critical for maximally activating the particular allophone /#ap/. However, under these conditions, it is still quite possible that the subjects would have sufficient information necessary to identify the initial vowel as /a/. This occurs in context-sensitive coding theory because a variety of allophones appropriate to the phoneme /a/ have been

activated more than the allophone appropriate for any other phoneme. When this happens, one is able to identify the phoneme, but not its immediate phonemic context. Clearly, with the break occurring in the middle of the feature density distribution for /api/, one has surely reduced the intelligibility of the /api/ allophone and also the /p/ phoneme (class of [p] allophones).

Again it should be emphasized that context-sensitive coding theory does not predict that it is impossible to recognize one phoneme without recognizing its immediately adjacent phonemes. This would be an absurd prediction in any event. Context-sensitive coding theory ought to predict a positive correlation between the recognizabilities of immediately adjacent phonemes, under many conditions. This positive correlation should occur primarily under conditions of rapidly articulated speech (normal speaking and hearing conditions). In the Lehiste and Shockey experiments, which was modelled after that of Öhman (1966), the initial and terminal vowels were quite prolonged. This provided extremely good steady-state formant cues for the recognizability for the initial and terminal vowel phonemes, no matter what was done to the transition to and from the intervocalic stop consonant. Under such conditions, because of the foregoing remarks regarding redundancy of features for the recognizability of any allophone or any class of allophones (phoneme), one would expect very low or even zero correlations between the recognizabilities of adjacent phonemes. Thus, more careful examination of correlations between the recognizabilities of adjacent phonemes in experiments such as that of Lehiste and Shockey would be rather inappropriate for the evaluation of context-sensitive coding theory. However, it should be emphasized that a repetition of the Lehiste and Shockey experiment (or, even better, of experiments involving somewhat longer phoneme sequences in nonsense utterances) at normal speaking rates ought to provide a strong test of the context-sensitive coding theory in speech perception.

Lehiste and Shockey found that subjects could not operate above chance in identifying a terminal vowel, given the initial vowel and the transition to the intervocalic consonant. Similarly, they found that subjects could not operate above chance in identifying the initial vowel, given the terminal vowel and the transition from the intervocalic consonant. These findings suggest that whether or not there is any overlap in the time of occurrence of the features for these non-adjacent vowel phonemes, the features appropriate for recognition of each vowel are not overlapping in their character. Thus, the context-conditioned variation that occurs in a transition to the intervocalic consonant from the initial vowel as a result of the nature of the terminal vowel does not provide a cue for the recognition of the terminal vowel. Analogous statements can be made for the case of the recognizability of the initial vowel as a function of the transition from the intervocalic consonant to the terminal vowel. If this is found to be generally true of non-adjacent phonemes, then one can strengthen the prediction of context-sensitive coding theory to the effect that, with nonsense utterances, only the recognizability of the immediately adjacent phonemes will be positively correlated. Recognizability of phonemes separated by one or more intervening phonemes should have zero correlation according to this formulation. Of course, such zero correlations between the recognizabilities of non-adjacent phonemes should only be found in nonsense utterances. To the extent that subjects can identify words from a subset of all the phonemes or allophones in the word, this will induce a positive correlation between the recognizabilities of any pair of phonemes in the word. This occurs because the subject knows the phonemic and allophonic constituents of the particular words in his lexicon.

To clearly understand the operation of context-sensitive coding in speech perception, one must note the many ways in which a context-sensitive coding theory of speech perception departs from the more fam-

iliar model of serial recognition of successive, non-overlapping, phoneme-sized segments. Since the features appropriate for the identification of each allophone overlap in time (and probably also in character), it is natural, and indeed necessary, to assure that all of the allophone detectors are operating in parallel. That is to say, the acoustic cues provide input simultaneously to all appropriate context-sensitive allophone representatives. It must be assumed that the temporally distributed input for any particular context-sensitive allophone representative can be summed up to some maximum period of time.

When all of the acoustic cues for a particular word such as "struck" have been received at the context-sensitve allophone level, there will be some distribution of degrees of activation imposed on all of the context-sensitive allophone representatives. For the word "struck" under conditions of high intelligibility, this would mean that the particular five allophone representatives shown in Fig. 3 would be maximally activated. These allophone representatives would be strongly associated with the representative of the word "struck" at the concept level, producing maximal activation of this word representative, rather than any other word representative. As noted before, during delivery of the features for the word "struck", heightened activation of the /$_{s}$t$_{r}$/ allophone representative ought to increase the activation of the sets of all /_s$_{t}$/ and /$_{t}$r_/ allophone representatives, since each allophone representative in these sets should be strongly associated to the /$_{s}$t$_{r}$/ allophone representative. Thus, the input to each allophone representative need not be assumed to be entirely from the lower auditory feature level, but could also be partly from association between allophone representatives at the segmental level. In addition, an activated allophone representative may be partially activating different concept representatives which, in turn, could produce associative feedback to the allophone representatives appropriate for those words.

It is difficult to make quantitative predictions from such a complex theory of the speech recognition process. However, it is clear that this type of system provides maximal use for speech recognition of the information in the acoustic signal.

Clearly, the theory is asserting that the speech recognition process is largely parallel, except for the fact that the acoustic cues for a word are to some extent spread out in time. It is the activation of a sufficient proportion of the unordered set of allophone representatives in a word that produces recognition of that word, not the activation of an ordered set of allophone, phoneme, or syllable representatives. Freeing the speech recognition process from the necessity of recognizing segments in temporal order greatly increases the power and flexibility of the speech recognition system. For one thing, it makes right-to-left effects possible in addition to left-to-right effects.

Recognizing the order of segments

It is obviously critical that we perceive the order of the phonemes within a word, in some manner, since the same phonemes in different orders often constitute different words. The context-sensitive coding theory asserts that the order of phonemes is represented by an unordered set of context-sensitive allophones (overlapping phoneme triples). In this theory, the context sensitivity of the successive allophone segments is the key to the representation of their order.

The context-sensitive coding theory of the representation of segmental order is directly supported by experiments such as that of Warren, Obusek, Farmer, and Warren (1969) which showed that human beings are extremely poor at recognizing the order of even an extremely short series of context-free elements such as hisses, buzzes and tones. Human beings apparently do not have much ability to represent the order of rapidly occurring events (durations of several 100

msec) unless such event sequences have occurred frequenlty in the past. Presumably frequent exposure to sequences of different events permits the establishment in the organism of units that represent something like overlapping triples of events (context-sensitive coding).

It is interesting to contrast the Warren, et al. findings with the findings of Yntema, Wozencraft, and Klem (1964) on short-term memory for lists of rapidly spoken digits. The lists in the Yntema, et al. study were random orderings of a set of eight digit names stored in a computer. The phonemes within the computer-spoken digits, of course, exhibit co-articulatory effects within the name for the digit. However, the transitions from one digit name to the next digit name in no way exhibit co-articulatory effects, since the same context-free digit name was used in all sequences and at all positions in the computer-spoken list of digits. Nevertheless, at rates where Warren, et al. found subjects completely unable to perceive the order of context-free hisses, buzzes, and tones, Yntema, et al. found subjects perfectly able to perceive the order of digits up to three or four digit lists (rates of two or four digits per second). Variations in rate in the Yntema, et al. study were achieved by introducing blank spaces between digit names. The digit names were always limited to a hundred milliseconds in length (which is a speeded-speech representation of each digit). Thus, intelligibility could certainly be improved if the auditory cues for each digit occupied the entire 250 msec at the four per second rate. This would undoubtedly have improved ordered memory span performance even more. Thus, there is a sharp contrast between the results for digit sequences and the results for sequences of hisses, buzzes, and tones. In the latter case, the order of even three or four such sounds could not be perceived above chance level at the rate of 250 msec per sound even when the sequence was repeated over and over again for as long as the subject desired before making his order judgement! Surely, if subjects listened to

a list of "context-free" digits over and over again they could achieve memory spans of at least seven or even more such digits.

From the standpoint of context-sensitive coding theory, one either has to assume: (a) that one can have context-sensitive coding of (multiphonemic) dif-it names, so that one has a representative for $/_2 5_4/$ which is different from the representation of a /5/ in any other context of (b) that the terminal and initial phonemes of each digit name are entering into context-sensitive allophone representation, without the context-conditioned variation of the transitions between them being present. In either case, there is no context-conditioned variation (co-articulatory cue) to signal the context-sensitively coded segments. Co-articulatory cues are very useful cues for context-sensitive segments, but they are by no means logically necessary for context-sensitive coding to work. Context-sensitive coding is most basically a theory of the representation of the ordering of segments in terms of temporally overlapping units. Context-sensitive coding of an ordered list of segments is possible even if the segments exhibit complete acoustic and articulatory invariance across all different contexts. Of course, such context-conditioned acoustic transitions constitute particularly good cues for the recognition of each context-sensitive unit (allophone or whatever). However, this does not prevent the recognition of a context-sensitive unit in the absence of such transitional cues.

Feedback from concept level to segment level in speech recognition

Experiments by Warren (1970) and Warren and Obusek (1971) provide important evidence for the existence of the previously postulated feedback from the concept (word) level to the segment level in speech recognition. In these experiments, a single phoneme, such as /s/, or an entire syllable, /gis/, was removed completely (including transitions to and

and from the segment) from the word "legislatures" and replaced by a cough, tone, or buzz. Subjects were not only able to correctly recognize the word, but also appeared to automatically fill-in the missing segment at a phonetic (allophonic) level. Subjects reported that no segment was missing, that they "heard" the missing phoneme(s). Furthermore, they were unable to judge accurately which segment was deleted, even when they were guaranteed that some segment had been deleted! Since subjects can recognize nonsense materials, we know that we are not denied conscious access to the phonetic (subconcept) level for perceptual judgements. Furthermore, under other conditions (when the gap was left as a silent interval not filled-in with any extraneous noise), subjects were able in the Warren experiments to accurately judge which segment was missing. Thus, the inability to judge the position of the missing segment under the initial set of conditions provides evidence that, under some conditions, the feedback from the concept level to the context-sensitive allophone level can, in conjuntion with random noise input, be sufficient to activate the missing context-sensitive allophone representatives. As mentioned previously, representation of a word in terms of an unordered set of context-sensitive segment representatives makes possible the efficient realization of this feedback from the concept to the segmental level, since the time at which a segment is activated is not important for the representation of the word at the segmental level.

Reaction time to segments of different size

The Savin and Bever (1970) experiment cited by Lehiste, and to a lesser extent the similar experiment of Warren (1971), provide additional evidence for the theory that the context-sensitive allophone is the basic unit of perception, rather than the phoneme. In the Savin and Bever experiment, subjects were to monitor a sequence of nonsense syllables for

either a single initial consonant phoneme /b/ or for an entire nonsense syllable that began with the phoneme /b/. Subjects responded more quickly to the syllable than to the initial /b/ phoneme in every case. The results were replicated with initial /s/ and for a medial vowel, /ae/. Of course, these results do not provide support for the context-sensitive allophone over the syllable as the minimal segmental unit for speech perception. What these results do suggest is that the minimal unit of speech perception is larger in temporal scope than the phoneme. Strong support for the context-sensitive allophone as the unit responsible for this effect, rather than the syllable, would come if the reaction time could be shown to decrease as one increased the phonemic size of the target from one to three phonemes (for syllables longer than 3 phonemes), but not to decrease for further increases in the phonemic length of the target (up to the length of the entire syllable).

References

Daniloff, R., and Moll, K. Coarticulation of lip rounding. Journal of Speech and Hearing Research, 1968, 11, 707-721.

Ladefoged, P. The phonetic framework of generative phonology. UCLA Working Papers in Phonetics No. 14, March, 1970, 25-32.

Ladefoged, P., and Silverstein, R.O. The interruptibility of speech. UCLA Working Papers in Phonetics No. 14, March, 1970, p. 10.

Lashley, K.S. The problem of serial order in behavior. In L.A. Jeffress (Ed.), Cerebral Mechanisms in Behavior. New York: Wiley, 1951.

Lehiste, I., and Shockey, L. The perception of coarticulation. Two papers presented at the 82nd meeting of the Acoustical Society of America, Denver, October 20, 1971.

MacKay, D.G. Forward and backward masking in motor systems. Kybernetik, 1969, 2, 57-64.

MacKay, D.G. Spoonerisms: The structure of errors in the serial order of speech. Neuropsychologia, 1970, 8, 323-350.

MacKay, D.G. Stress pre-entry in motor systems. American Journal of Psychology, 1971, 84, 35-51.

Öhman, S.E.G. Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America, 1966, 39, 151-168.

Savin, H.B., and Bever, T.G. The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 295-302.

Warren, R.M. Perceptual restoration of missing speech sounds. Science, 1970, 167, 392-393.

Warren, R.M. Identification times for phonemic components of graded complexity and for spelling of speech. Perception and Psychophysics, 1971, 9, 345-349.

Warren, R.M., and Obusek, C.J. Speech perception and phonemic restorations. Perception and Psychophysics, 1971, 9, 358-363.

Warren, R.M., Obusek, C.H., Farmer, R.H., and Warren, R.P. Auditory sequence: Confusion of patterns other than speech or music. Science, 1969, 164, 586-587.

Wickelgren, W.A. Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychological Review, 1969(a), 76, 1-15.

Wickelgren, W.A. Context-sensitive coding in speech recognition, articulation, and development. In K.N. Leibovic (Ed.), Information Processing in the Nervous System. New York: Springer-Verlag, 1969, 85-95.

Yntema, D.B., Wozencraft, F.T., and Klem. L. Immediate serial recall of digits presented at very high rates. Presented at the meeting of the Psychonomic Society, Niagara Falls, Ontario, October, 1964.

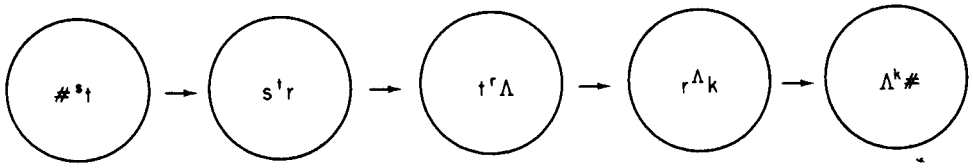

Fig. 1. Associative chain of context-sensitive allophone representatives for the word "struck."

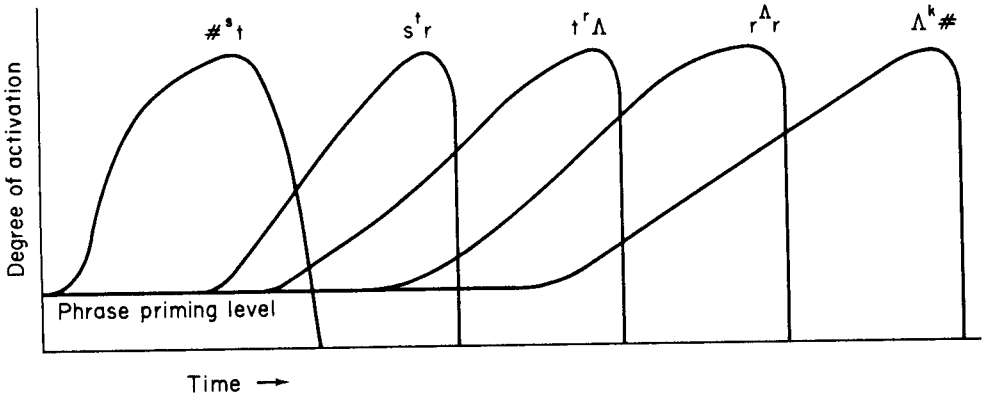

Fig. 2. Hypothetical approximate degree of activation of context-sensitive allophone representatives in articulation of the word "struck." Note that, in general, one cannot assume that all allophone representatives have approximately equal duration of maximal activation at any given rate of talking.

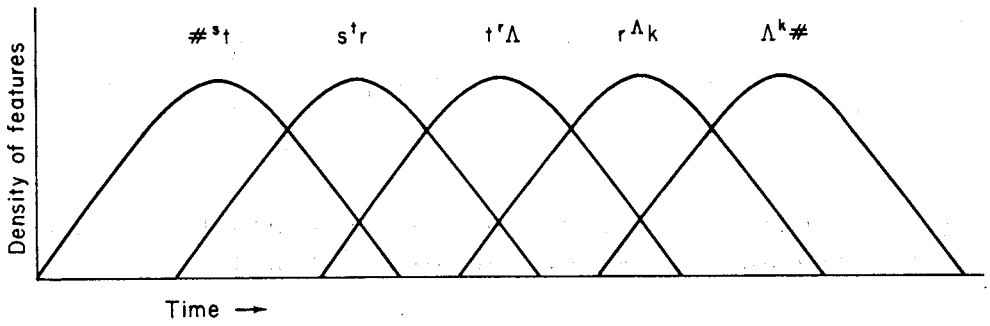

Fig. 3. Approximate density of features for recognition of each context-sensitive allophone in the word "struck."

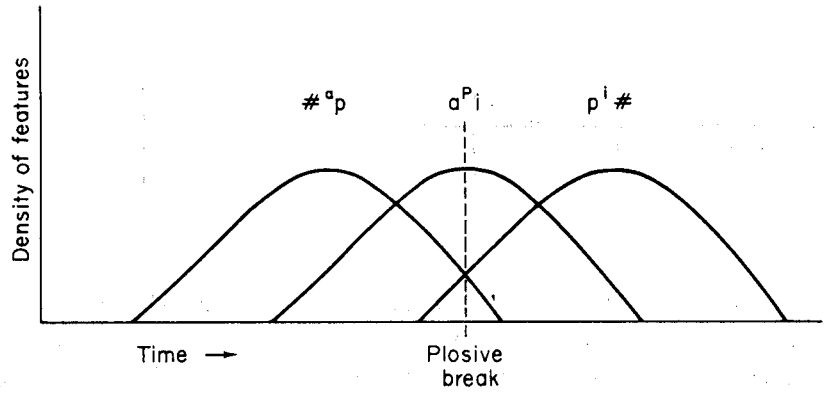

Fig. 4. Approximate density of features for recognition of each context-sensitive allophone in the nonsense word "api."